# Week 5 Summary: Apriori, FP-Growth, SON, and Pattern Evaluation

Z3003 Data Mining: Scaling to Large Data · 14–16 October 2026 · Dr. Innocent Nyalala

## Lecture 13: Frequent Itemsets and the Apriori Algorithm

Mining proceeds in two stages: (1) find all itemsets with support $\geq$ min_sup, then (2) emit rules with confidence $\geq$ min_conf. By the anti-monotone Apriori principle ($A \subseteq B \implies \text{support}(A) \geq \text{support}(B)$), every subset of a frequent itemset is frequent, so any candidate $C_k$ containing an infrequent $(k-1)$-subset is pruned without reading the dataset. On $120,000$ agro-dealer baskets, Apriori discovers 60 frequent itemsets while counting roughly 250 candidates instead of $4.19 \times 10^6$.

## Lecture 14: Scaling Pattern Mining (FP-Growth, PCY, Sampling, SON)

| Algorithm | Disk Passes | Candidate Generation | Core Mechanism |
|---|---|---|---|
| Apriori | $k_{\max}$ passes | Explicit $C_k$ | Anti-monotone subset pruning |
| FP-Growth | 2 passes | Zero candidates | Prefix FP-tree + conditional pattern bases |
| PCY | 2 passes (for pairs) | Bucket-filtered $C_2$ | Hash pair counts during pass 1 |
| Random Sampling | 1–2 passes | Approximate | Scale threshold on in-memory sample |
| SON | 2 passes | Chunk union | Exact parallel mining on MapReduce/Spark |

## Lecture 15: Null-Invariant Pattern Evaluation

Because null transactions (containing neither $A$ nor $B$) inflate lift and $\chi^2$ on sparse datasets, pair **Kulczynski** $\text{Kulc}(A,B) = \frac{1}{2}\big(P(B \mid A) + P(A \mid B)\big)$ with **Imbalance Ratio** $\text{IR} = \frac{|c(A)-c(B)|}{c(A)+c(B)-c(A \cup B)}$ to rank genuine symmetric rules.

**Worked Check: Null-Transaction Effect**

With $10,000$ baskets (milk $6,000$, bread $1,000$, both 800), lift $= 1.33$, Kulc $= \frac{1}{2}(0.80 + 0.133) = 0.467$, and IR $= 0.81$. Adding $90,000$ empty baskets inflates lift tenfold to 13.3 yet leaves Kulc $= 0.467$ unchanged.